



OPEN ACCESS

EDITED BY

Valentina Tozzini,
National Research Council (CNR), Italy

REVIEWED BY

Shivansh Kaushik,
University of California,
Riverside, United States
Alessandro Mossa,
National Research Council (CNR), Italy

*CORRESPONDENCE

Morten Nielsen
✉ morni@dtu.dk

RECEIVED 22 April 2025

ACCEPTED 18 June 2025

PUBLISHED 17 July 2025

CITATION

Deleuran SN and Nielsen M (2025) NetTCR-
struc, a structure driven approach for
prediction of TCR-pMHC interactions.
Front. Immunol. 16:1616328.
doi: 10.3389/fimmu.2025.1616328

COPYRIGHT

© 2025 Deleuran and Nielsen. This is an open-
access article distributed under the terms of
the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

NetTCR-struc, a structure driven approach for prediction of TCR-pMHC interactions

Sebastian N. Deleuran and Morten Nielsen*

Department of Health Technology, Technical University of Denmark, Lyngby, Denmark

Accurate modeling of T cell receptor (TCR)–peptide–major histocompatibility complex (pMHC) interactions is critical for understanding immune recognition. In this study, we present advances in structural modeling of TCR-pMHC class I complexes focusing on improving docking quality scoring and structural model selection using graph neural networks (GNN). We find that AlphaFold-Multimer's confidence score in certain cases correlates poorly with DockQ quality scores, leading to overestimation of model accuracy. Our proposed GNN solution achieves a 25% increase in Spearman's correlation between predicted quality and DockQ (from 0.681 to 0.855) and improves docking candidate ranking. Additionally, the GNN completely avoids selection of failed structures. Additionally, we assess the ability of our models to distinguish binding from non-binding TCR-pMHC interactions based on their predicted quality. Here, we demonstrate that our proposed model, particularly for high-quality structural models, is capable of discriminating between binding and non-binding complexes in a zero-shot setting. However, our findings also underlined that the structural pipeline struggled to generate sufficiently accurate TCR-pMHC models for reliable binding classification, highlighting the need for further improvements in modeling accuracy.

KEYWORDS

T cell receptor, protein structure prediction, docking, TCR specificity prediction, machine learning

Introduction

T cells drive the adaptive immune response by recognizing and eliminating cells displaying foreign peptides through major histocompatibility complexes (MHC) (1). This process is facilitated by interaction between the T cell receptor (TCR) and the peptide-MHC complex (pMHC) which serves as a crucial checkpoint for immune activation. Understanding the rules governing this interaction is critical, as it is central to the development of a wide range of immunotherapy treatments.

Computational prediction of this event presents an effective avenue of greatly accelerating development of immunotherapies, and consequently a wide range of methods aiming to achieve this have been developed (2–6). Primarily, these methods represent the TCR and pMHC using their amino acid sequences, while often utilizing a reduced representation of the TCR by only considering either all CDRs or only the most variable and specificity-defining CDR3 (7, 8). While some success in the many-shot learning setting has been demonstrated, the zero-shot setting, i.e. inference on completely unseen TCRs and peptides, remains largely unsolved (9). An estimate of 10^8 unique TCR β sequences may exist in a single individual that, through cross-reactivity and alternative binding modes, may interact with the staggeringly high number of 20^9 9-mer amino acid combinations (10, 11). Given that approximately only 50,000 paired chain TCR-pHLA class I interactions, not accounting for redundancy, have been described in major databases IEDB and VDJdb, the poor zero-shot performance observed in current state-of-the-art models is unsurprising (2, 12, 13).

Recent advances in protein structure prediction have allowed accurate structural modeling of TCR-pMHC complexes (14–16). This provides a new avenue in tackling the TCR specificity prediction task, as the conserved nature of protein structure may work as a significantly less diverse perspective on TCRs and peptides.

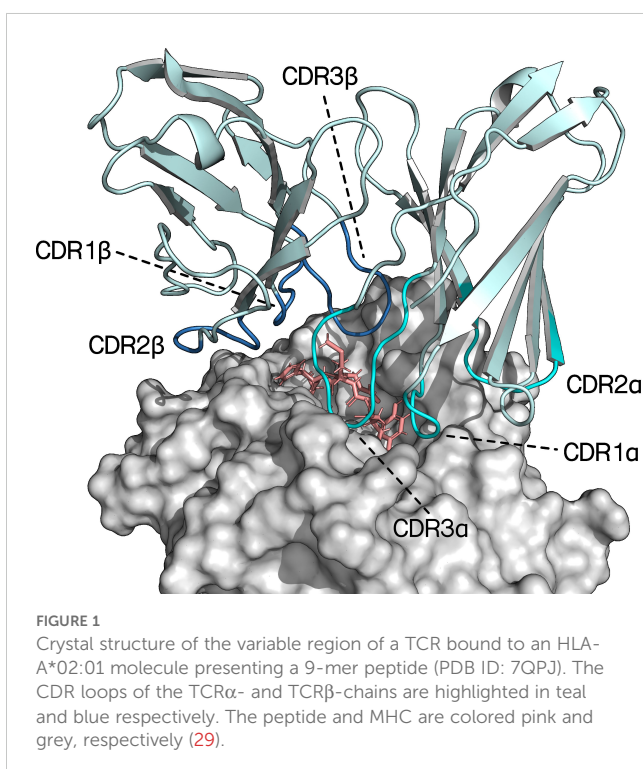
However, at this time, few methods explicitly utilize structural data and those that do, have not demonstrated significant gains in performance over methods trained on sequence data (17). Besides the computational cost of performing large scale experiments with state-of-the-art protein modeling tools such as AlphaFold (AF), additional challenges in utilizing structural data are present.

Accurate modeling of TCRs, especially in their docked conformations, is an immensely challenging task due to their highly variable and long CDR3 loops (Figure 1). Further analysis of these structural models also remains difficult, given that stimulatory TCR binding can depend on the formation of very few contacts that may not be captured even in high quality models (18). Recent work by Motmaen et al. and Yin et al. on utilizing AF for modeling pMHC and TCR-pMHC complexes by modifying pMHC and TCR multiple sequence alignment (MSA) and template featurization, have provided significant improvements in modeling accuracy over AlphaFold-M (AF-M) (15, 19). The latter approach demonstrates high modeling accuracy on a benchmark dataset of solved TCR-pMHC structures. However, the confidence metric provided by AF-M appears to not correlate strongly with the modeling quality of TCR-pMHC interfaces, when quantified with the DockQ metric (20). Consequently, selection of high quality models from the pool of modeled structures is not guaranteed. Here, we propose a graph neural network based approach for significantly more accurate quality scoring of TCR-pMHC complexes modeled using AF-M. The model is based on geometric vector perceptron layers allowing for fine grained encoding of geometric features, and is trained to perform regression on the DockQ docking quality metric (21). In order to generate training data for this model, we additionally present results for perturbation of AF MSA and template features and its effect on modeling diversity and quality. Secondly, we apply our docking quality scoring approach to the task of TCR-pMHC binding prediction and demonstrate its efficacy in correctly ranking binding TCR-pMHC complexes when sufficiently accurate structural models are available. While the pipeline in principle could be applied to any receptor binding prediction task, we have in this work specifically focused on the TCR-pMHC binding prediction task with the aim to address the unsolved challenge of predicting specificity of T cell receptors towards unseen pMHC complexes. This is reflected in the domain specific features of the proposed pipeline.

Materials and methods

Modeling pipeline

Structures of TCR-pMHC complexes were modeled using an AF-Multimer version 2.3 based pipeline. Multiple sequence alignment (MSA) and template features were created using the approach described by Yin et al. (15). Here, template features for the pMHC are generated such that the pMHC is modeled as a single chain, allowing for the use of docked pMHC templates. Additionally, TCR MSA and template features are generated from a reduced database of immunoglobulin proteins. In order to increase modeling throughput on high performance compute clusters, the featurization and modeling steps of AF were decoupled so that featurization for a batch of sequences could be run in parallel with modeling once features for the first entry in the batch is completed.



An additional set of options for perturbing MSA and template features were added to the pipeline, with the aim of increasing diversity of the set of structures predicted for a given input (22). These include random mutation in the MSA, column wise mutation in the MSA, masking of MSA hits (resembling MSA subsampling) and addition of gaussian noise to structural template atomic coordinates. An option to enable dropout of AF modules was also added.

Training dataset for docking quality scoring

The dataset used for training and evaluating the GVP-GNN DockQ regressor was constructed from a set of solved TCR-pMHC class I complex structures. Structures were obtained from RCSB and TCRs were trimmed to their variable domains (23). Complexes that contained peptides with non-standard amino acids were removed from the dataset. Additionally, structures were filtered to human complexes with $\alpha:\beta$ TCRs. Finally, a resolution cutoff of 3.5Å was applied resulting in a final data set of 80 structures.

For the cross-validation setup, partitions were created in the following manner. Structures released after the AF-M 2.3 training dataset cutoff of 2021-09-30 were selected for use as a benchmark dataset. The Hobohm 1 algorithm was applied to the data for redundancy reduction, using a 95% sequence similarity threshold. Sequence similarity was calculated over the alignment length, and any complex with a TCR α or TCR β sequence that was 95% similar to an already encountered sequence was dropped. Following this, for the training data structures, complete linkage agglomerative clustering based on TCR α or TCR β sequence similarity was applied in order to generate 5 partitions. Here, the AgglomerativeClustering method from scikit-learn was applied to a matrix of pairwise average TCR α or TCR β sequence similarities, setting the number of desired clusters to 5 and otherwise using the default settings provided in scikit-learn v1.0.2. The resulting clusters were then used to define the partitions containing 19, 18, 6, 6 and 6 solved structures. The benchmark dataset contained 25 solved structures.

Sequences from the solved structures were extracted and used to generate structural models using the AF-M based pipeline. In order to increase the diversity of modeling quality for each target in the dataset, the input feature perturbation methods described in the Modeling pipeline section were applied on MSA and template features for each modeling seed. Structures were modeled under different pipeline configurations in order to increase the uniformity of the docking pose DockQ distribution. The following runs were performed in order to achieve this:

- No restriction on template selection, 30 candidates per AF model (150 total), and a maximum number of recycling of 3.
- No template information, except for the pMHC, masking of 60% of MSAs, random substitution of 60% of MSA residues, 30 candidates per AF model (150 total).
- No template information, except for the pMHC, masking of 20% of MSAs, random substitution of 20% of MSA residues, 30 candidates per AF mode (150 total).
- Maximum template date set to AF-M 2.3 training dataset cutoff (2021-09-30) and a template and query sequence similarity threshold of 90%, masking of 15% of MSAs, random substitution of 15% of MSA residues, 60 candidates per AF model (300 total).

This pipeline, thus resulted in 750 candidate structural models being generated for each input TCR-pMHC entry. Subsequently models were scored against their ground truth targets using DockQ, considering only the TCR and peptide interactions. As a measure against redundancy, for each target, models from the training/validation partitions were placed in 20 bins according to their DockQ score and up to 20 models from each bin were sampled. Additionally, a second dataset was created where only models with DockQ ≥ 0.5 were retained, and where no redundancy reduction was performed w.r.t. the DockQ distribution. Finally, for targets in the benchmark dataset, models with more than 5 TRA-peptide or TRB-peptide backbone clashes were filtered. A backbone clash was defined as backbone atoms in peptide and TCR being within 3Å of each other. This resulted in two training datasets with 12057 (full DockQ range) and 16541 (DockQ ≥ 0.5 and no homology reduction) models over 55 targets and a benchmark set of 3750 models for 25 targets.

The GVP-GNN regressor

DockQ regression was performed using a geometric vector perceptron graph neural network (GVP-GNN) (21) (Figure 2). The graph representation of a given protein complex was generated in a per-residue manner, where nodes represented amino acid residues and edges were created based on the euclidean distances of these nodes. As a regularizing measure, the model was trained using a pair mean squared error (MSE) loss function, as described by Jing et al. (21):

$$Loss = MSE(Y1_{true}, Y1_{pred}) + MSE(Y2_{true}, Y2_{pred}) + MSE(Y1_{true} - Y2_{true}, Y1_{pred} - Y2_{pred})$$

Where $Y1_{true}$, $Y1_{pred}$, $Y2_{true}$, $Y2_{pred}$ are predicted and true DockQ values for the same target structure Y. The difference term, $MSE(Y1_{true} - Y2_{true}, Y1_{pred} - Y2_{pred})$, aims to reduce fitting to features that are static between candidates of the same target such as amino acid sequence, by penalizing cases where one candidate is scored more accurately than another candidate, potentially caused by overfitting to these static features.

Models were trained in a 5-fold cross-validation setup to obtain an ensemble of models. We trained two sets of models, one on a dataset composed only of structural models with a DockQ score over 0.5, and one trained on a dataset spanning the full range of the DockQ metric. This resulted in a set of 10 models. Models were constructed and trained with the following hyperparameters:

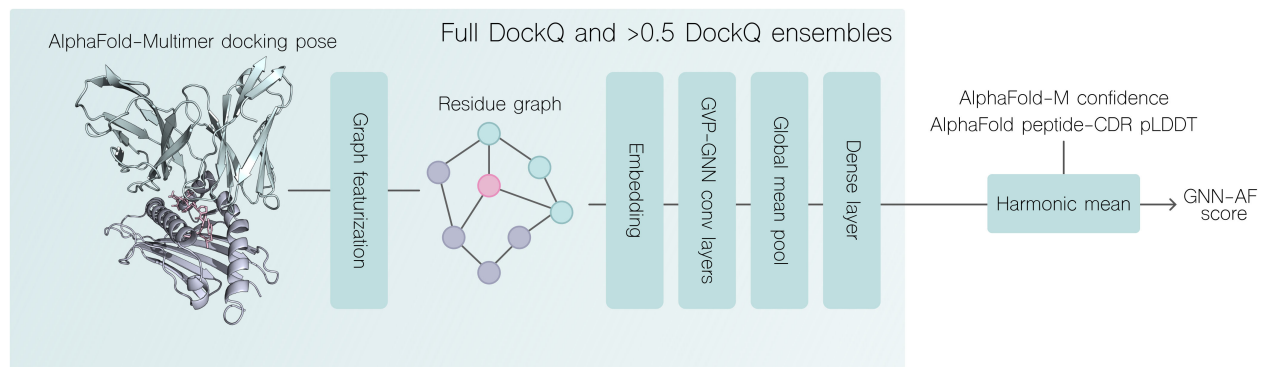


FIGURE 2

Overview of DockQ regression and quality prediction pipeline. A docking pose generated with AlphaFold-Multimer is featurized as a graph (see materials and methods). Two sets of 5 GNN-GVP models resulting from training on 5 cross-validation splits, where one set was trained only on docking poses with DockQ >0.5 are used to score the structural model. The combined GNN-AF score is computed by the harmonic mean of the two GVP-GNN ensemble scores, AF-M confidence and the peptide-CDR pLDDT score. For more details, see materials and methods.

- Optimizer: Adam
- Initial learning rate: 0.0001
- Learning rate scheduler: Cosine annealing (CosineAnnealingLR)
- Weight decay (L2 regularization): 0.0001
- Batch size: 32
- Number of training epochs: 100
- Dropout rate: 0.5 (applied to fully connected layers)
- Graph convolution layers: 3
- Node hidden dimension (scalar features): 32
- Node hidden dimension (vector features): 8
- Edge hidden dimension (scalar features): 50
- Edge hidden dimension (vector features): 8

We define a GNN ensemble score for a given structure, as the harmonic mean of two predicted DockQ values, which are obtained from taking the arithmetic mean of the 5 GNNs trained on the full DockQ range, and the arithmetic mean of the 5 GNNs trained only on structural models with a >0.5 DockQ.

We additionally define a consensus predicted quality score for a given input as the harmonic mean of the two GNN scores, AF-M confidence metric and CDR-peptide pLDDT. The CDR-peptide pLDDT was computed by extracting the pLDDT scores associated with CDR123αβ and peptide residues, and taking the arithmetic mean of these scores. We term this score GNN-AF. Thus, the predicted quality is given by:

$$GNN_AF = \text{Harmonic mean}(GNN_{ensemble_Full_DockQ}, GNN_{ensemble_>0.5_DockQ}, AF_{confidence}, AF_{CDRpep_pLDDT})$$

We chose to use the harmonic mean, as it will give more weight to smaller numbers when computing the mean. This is useful, as the AF confidence metric is more prone to overestimation rather than underestimation of quality, and because the graph neural networks developed here can predict DockQ values in the lower range

accurately. Thus, when either of these scores are low for a given input, we want to ensure that this input is assigned a low score.

Data featurization

Graph featurization of protein structures was performed following the methods described in Jing et al. (21). Here, each node corresponded to a residue in the full protein complex. For each node v_i , its set of edges was defined as those connecting v_i to its 30 nearest neighbors measured by euclidean distance. Scalar and vector node features were defined as suggested by Jing et al. (21). Let v_i be a node representing an amino acid in the i 'th position of the concatenated sequence:

- Scalar features residue v_i :
 - $\{\sin, \cos\} \circ \{\phi, \psi, \omega\}$, where ϕ, ψ, ω are the dihedral angles of v_i .
 - One hot encoding of the amino acid of v_i .
 - One hot encoding of the chain that v_i belongs to.
- Vector features of residue v_i :
 - Unit vectors describing the relative orientation of v_{i-1} and v_{i+1} w.r.t to their $C\alpha$ -atoms.
 - Unit vector describing the imputed direction of $C\beta_i - C\alpha_i$, computed by:

$$C\beta_i - C\alpha_i = \sqrt{\frac{1}{3}} \frac{(n \times c)}{\|n \times c\|_2} - \sqrt{\frac{2}{3}} \frac{(n + c)}{\|n + c\|_2}$$

where $n = N_i - C\alpha_i$ and $c = C_i - C\alpha_i$.

Let e_{ij} be a directed edge between two nodes v_i and v_j .

- Scalar features of e_{ij} :
 - An encoding of $\|C\alpha_j - C\alpha_i\|^2$ in terms of 16 Gaussian radial basis functions.

- A sinusoidal positional embedding representing the distance of v_i and v_j in the primary structure. For inter-chain edges, the positional embeddings were set to 0, as the distance in primary structure is not defined here.
- Vector features of e_{ji} :
 - Unit vector describing the direction $C\alpha_j - C\alpha_i$.

We additionally present a GNN model, where we incorporate structural embeddings as additional node scalar features. These embeddings are extracted from the final layer of the encoder module of the ESM-IF1 inverse folding model (24). Briefly, this model is trained to generate amino acid sequences that are likely to fold into the geometry specified by an input protein structure. The model uses a GVP-GNN module to generate structural features, followed by a generic encoder-decoder transformer module. Each embedding has a shape of $512 \times N_{AA}$ where N_{AA} represents the total length of all chains in a TCR-pMHC complex. For each residue i , its corresponding embedding vector was concatenated with the scalar node feature vector for v_i . We denote this model GNN-ens-IF1, and GNN-IF1 when used in the consensus score.

TCR specificity data

The binding classification dataset was derived from Jensen (2). Briefly, the positive examples in this dataset were derived from IEDB, VDJdb as well as a 10x Genomics sequencing dataset denoised using the ITRAP algorithm (12, 13, 25, 26). Swapped negatives were generated by mismatching each positive TCR-pMHC pair with 5 TCRs positive to other pMHCs. In this study, we opted to downsample the dataset with up to 200 positive examples per peptide and resample negative examples following the same procedure as described by the authors. Partitions were then generated from this downsampled dataset, again using the procedure described by the authors. For more details, we refer to the methods section of Jensen (2). The downsampled dataset was composed of 2945 binding complexes and 14725 (5×2945) swapped complexes giving a total of 17670 examples.

Modeling of TCR-pMHC binding dataset

For each set of sequences in the retrieved TCR-pMHC binding dataset, 10 candidates for each of the 5 AlphaFold-M multimer models were modeled in total using a 90% sequence similarity threshold on TCR α , TCR β and peptide template selection, in order to prevent entries with high similarity to existing solved structures being more accurately modeled. There was no restriction on MHC template sequence similarity, as it was assumed that structural variability for all relevant MHC molecules was sufficiently described.

Batch sampling experiment

For each peptide, binding and non-binding complexes were partitioned into batches of 1 binder and 5 non-binders, such that the batch contains complexes all with the same pMHC, for which one TCR is positive and the remaining TCRs negative. Given a batch, a scoring method is then tasked to assign a score to each of the 6 complexes, such that the positive binding complex is ranked as the highest scoring complex. Performance metrics are then computed in a per batch manner, evaluating how often binding complexes are correctly ranked. Here, we use a true-positive rank metric we term TPR that expresses how many non-binding complexes are scored higher than the binding complex:

$$TPR = \frac{batch_length - 1 - binder_index}{batch_length - 1}$$

Additionally, we compute an additional metric we term batch accuracy. Here, a batch is considered correctly predicted if the binding complex is ranked as number one ($binder_index == 0$), and incorrectly predicted if not. The batch accuracy is then computed across all batches indicating the proportion of batches where the positive example was assigned the highest score (Supplementary 4).

Results

In order to effectively use structural models for predicting TCR-pMHC binding, we first set out to increase modeling quality by improving on the docking candidate selection step in AlphaFold. For this purpose, we developed a graph neural network based DockQ regressor, trained on a set of modeled TCR-pMHC complexes. Subsequently, we retrieved a large set of TCR sequences annotated with their peptide specificity from public databases and datasets, generated non-binding complexes by mismatching TCRs and pMHCs, and used our newly developed modeling pipeline to generate a set of binding and non-binding TCR-pMHC complex structural models. Using these structures, we then applied a range of deep learning based methods in order to classify binding and non-binding complexes.

Structural diversity of AlphaFold-Multimer models

Initial experiments with modeling TCR-pMHC complexes using AlphaFold-Multimer (AF-M), revealed that for some benchmark dataset targets, many candidates would assume the same, incorrect configuration of CDR loops. This in turn led to loss of important native contacts of CDR3 loops, that would presumably hinder useful inference about the modeled structures and their immunological properties. We note that, we here use the benchmark dataset as a validation set for the development of the modeling pipeline, in order

to quantify its capabilities for modeling the large set of sequences in the TCR specificity dataset. Inspired by Stein and Mchaourab, we developed a set of tools to increase stochasticity in the AlphaFold-Multimer featurization pipeline and consequently reduce homogeneity of modeled candidates (22). A set of candidate structures was then modeled for targets in the benchmark dataset, using a combination of different perturbation configurations and the effect of the perturbation was quantified w.r.t. structural diversity using the median intra-target pairwise backbone RMSD and w.r.t. modeling quality quantified by the DockQ metric, for details refer to materials and methods (Figure 3).

When not applying MSA and template feature perturbation, the median of the per target median of median pairwise TCR RMSD is 1.47Å, indicating low structural diversity. This is especially pronounced for targets such as 7L1D, 7NDQ, 7RK7, where no or almost no candidate structures exceed a median pairwise RMSD of 4Å. Applying the feature perturbation method, the median of the per target median of median pairwise TCR RMSD values was increased to 6.12Å, showcasing a substantial increase in intra-target modeling diversity. However, this does not translate to an overall improvement in modeling accuracy, except for targets 7PB2, 7RK7 and 7RRG where we observe a max DockQ increase of approximately 0.1. Oppositely, other targets such as 7RM4, 7N2R

and 7DZN have a comparative decrease in max DockQ. Additionally, we see a poorer density of high quality candidates which will make selection of high quality candidates harder, as the median of the median per target DockQ is brought down from 0.63 to 0.45. Based on these results, we opted not to use feature perturbation as a default feature of the pipeline and only use this approach for generating training data for the scoring function optimization.

GVP-GNN regression performance and improvements over AlphaFold

Subsequently, we moved on to address the second challenge we encountered when modeling TCR-pMHC complexes using AF-M, namely that the confidence metric used internally in AF-M overestimates the quality for certain targets and consequently only correlates moderately with DockQ scores of candidates modeled for the benchmark dataset (see Figure 4a). This figure demonstrates that a large number of complexes with AF-M quality scores above 0.7 have DockQ scores below 0.23 [corresponding to CAPRI threshold for incorrect structures (20)]. We can further quantify this by calculating the cumulative proportion of incorrect

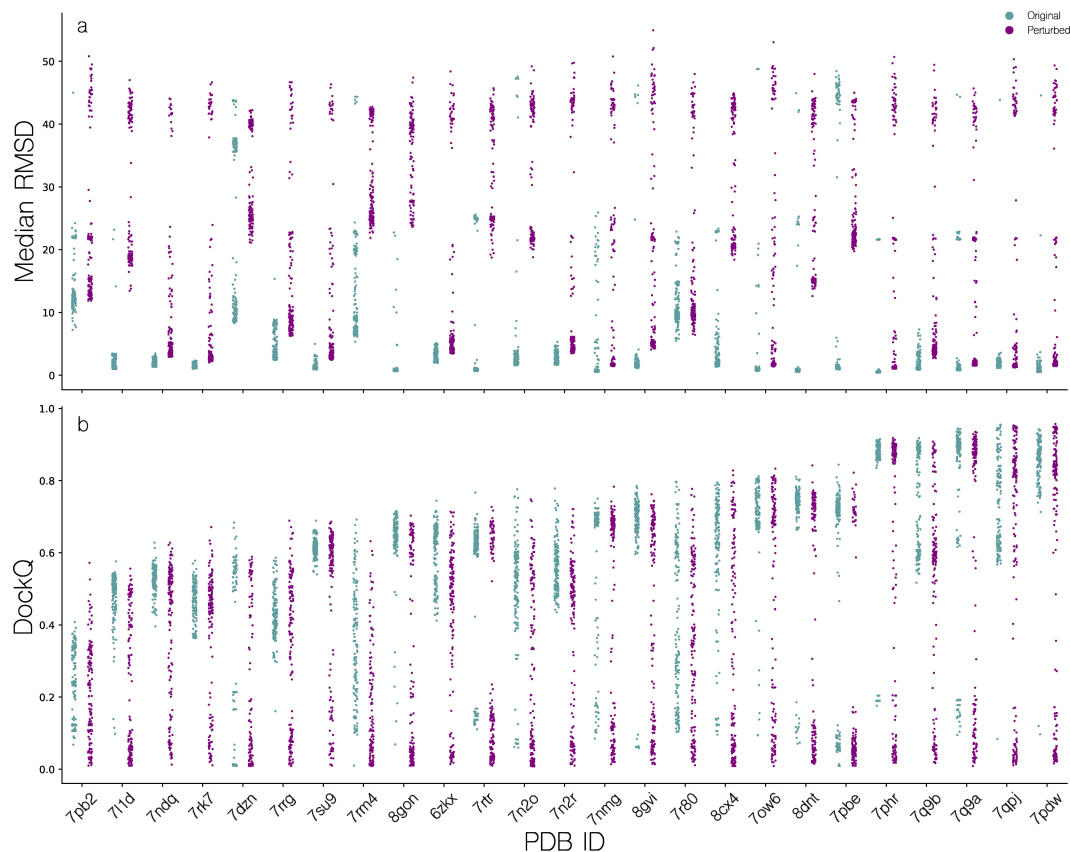


FIGURE 3

Distribution of quality scores for candidate structures modeled using AlphaFold-Multimer. (a) Median pairwise TCR RMSD (superimposed on pMHCs), for all candidates for a given target, before and after feature perturbation. (b) Distribution of DockQ scores for all candidates of a given target, before and after MSA and template feature perturbation.

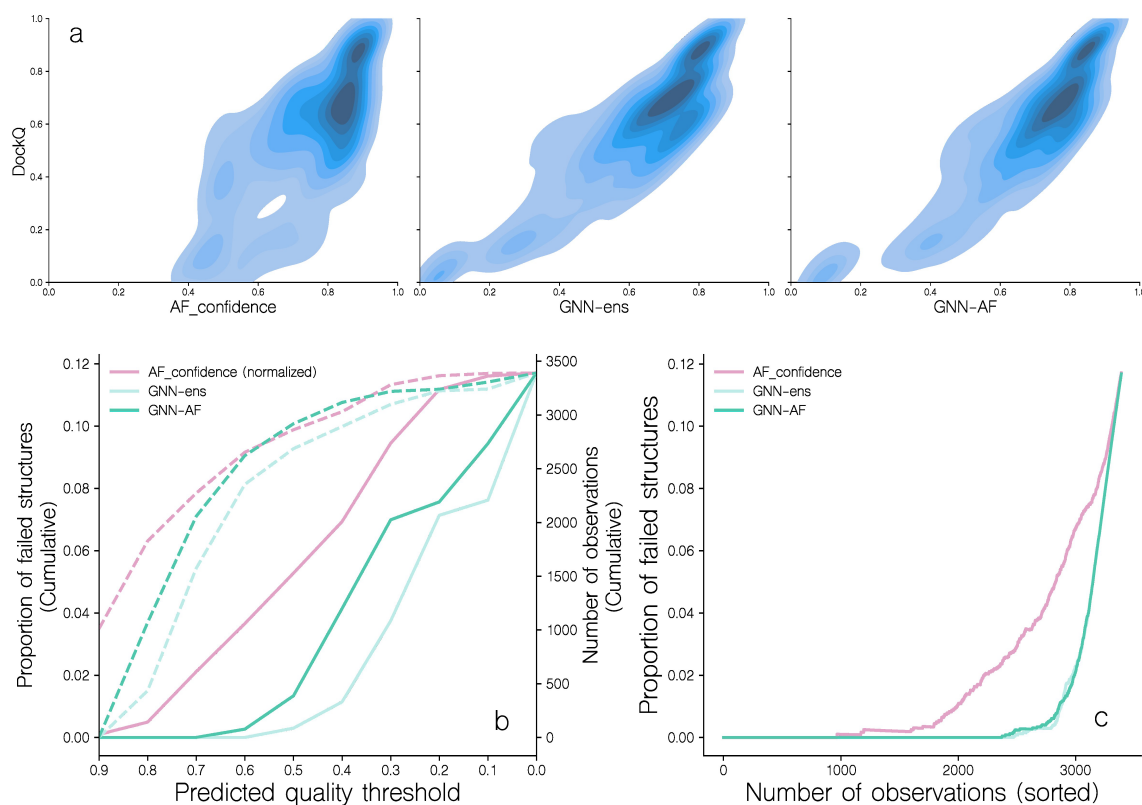


FIGURE 4

Distribution of DockQ scores for candidate structures ranked by different predicted quality scores. (a) Correlation of DockQ and AF_confidence, GNN-ens and GNN-AF for all candidates in the benchmark dataset. For more information on these scoring schemes, see the materials and methods section. (b) The proportion of failed candidates in 10 bins of either AF_confidence (normalized), GNN-ens, or GNN-AF score. Dotted lines indicate the number of observations in each bin. Following the CAPRI classification scheme, a candidate is considered as failed when DockQ < 0.23. (c) The proportion of failed candidates as a function of the number of candidates, when sorted by either AF_confidence, GNN-ens or GNN-AF scores.

structures (DockQ score less than 0.23) as a function of the annotated quality score (Figure 4b). Doing this, we find that AF-M, at a predicted quality score of ~0.7, has 2% of cumulative incorrect structures, within the set of ~2200 top scoring complexes (Figure 4c).

To address this issue, we developed a series of graph neural network based scoring methods to predict DockQ for a given input structure. Two GVP-GNN models were constructed, each trained in a 5 fold cross-validation setup, with data partitioned by reducing interpartitional joint TRA and TRB sequence similarity (for details on the redundancy reduction scheme see materials and methods. For a low dimensional visualization of structural features of the dataset refer to Supplementary 1). One was trained on the complete data set, and the other was trained only on docking poses with >0.5 DockQ (for details on the models and model training refer to materials and methods. For details on the models' training dynamics refer to Supplementary 2). The outputs from these two ensembles of 5 models were combined to predict DockQ for a given docking pose. We denote this score GNN-ens. Additionally, the outputs of these models were combined into a consensus score we denote GNN-AF, by computing the harmonic mean of the two GNN ensemble scores, AF_confidence and an additional AlphaFold score, AF_CDRpep_pLDDT (the mean of pLDDT scores associated

with CDR123 $\alpha\beta$ and peptide residues, for more information on this score, refer to materials and methods).

Next, the performance of these quality assessment scores was evaluated on the validation and benchmark datasets in terms of different correlation metrics between the predicted quality and measured DockQ values (see Table 1). This analysis demonstrated an overall improved performance of the GNN-based models compared to AF-M. For instance, the overall Spearman's rank correlation across all modeled structures improved from 0.681 to 0.824 (21%) for the GNN-ens model when compared to the AF-M confidence score. Notably, the correlation was greatly improved across the full DockQ range which in large parts resolves the issue of the AF-M confidence score overestimating the quality of poor quality candidates (Figure 4b). Analyzing the rank correlation for candidates within each TCR-pMHC target also revealed a major increase of 52.02% (0.454 versus 0.297) increase in mean Spearman correlation (refer to Supplementary 3 for more information on the scoring of individual targets). Together this indicates that the GNN-ens method ranks candidates globally and locally better than the AlphaFold confidence metric. The performance gain compared to using AF-M score alone was further improved when considering the GNN-AF-ens model, merging the GNN ensemble with the AF-M confidence and AF-peptide-CDR pLDDT scores.

TABLE 1 Scoring performance of GNN, GNN-IF1, AlphaFold and ensemble methods for the validation and benchmark data set.

Method	Validation				Benchmark			
	Global SCC	Global >0.5 DockQ SCC	Mean local SCC	Mean top 1 DockQ	Global SCC	Global >0.5 DockQ SCC	Mean local SCC	Mean Top 1 DockQ
AF_confidence	0.777	0.528	0.788	0.583	0.681	0.521	0.297	0.615
AF_CDRpep_pLDDT	0.751	0.576	0.745	0.590	0.750	0.681	0.264	0.599
GNN-ens	0.855	0.412	0.852	0.627	0.824	0.650	0.454	0.690
GNN-AF	0.877	0.539	0.864	0.646	0.855	0.710	0.502	0.673
GNN-IF1-ens	0.864	0.345	0.855	0.647	0.799	0.673	0.447	0.650
GNN-IF1-AF	0.891	0.460	0.869	0.642	0.818	0.688	0.459	0.643

Performance metrics are, Global SCC: Spearman's rank correlation coefficient (SCC) between predicted quality and DockQ, computed across all candidates for all targets, Global >0.5 DockQ SCC: SCC between predicted quality and DockQ, computed across all candidates for all targets with DockQ > 0.5, Mean local SCC: Mean per-target SCC between predicted quality and DockQ and Mean top 1 DockQ: Mean DockQ of the top 1 selection for each target. The GNN method is defined from the harmonic mean of the predicted DockQ score for two GVP-GNN ensembles. The GNN-IF1 method is defined from the harmonic mean of the predicted DockQ score for two GVP-GNN-IF1 ensembles. The GNN-AF and GNN-IF1-AF models are defined from the harmonic mean of the two GVP-GNN ensemble scores (for GVP-GNN and GVP-GNN-IF1 ensembles respectively), AF-M confidence and the AF CDR-peptide pLDDT scores. For details, see materials and methods. Bold numbers indicate the best score for each metric within each dataset (validation or benchmark).

Including the GNN-ens and GNN-AF-ens models in the analysis shown in Figure 4 further supports this gain in accuracy and demonstrates that at the score threshold where the proportion of failed structures reaches 2% for both methods correspond to ~3100 complexes, an increase of 25% compared to the number of structures captured at the equivalent proportion of failed structures of AF-M.

The top 1 candidate selection is also notably improved (see Figure 5; Table 1). Here, for almost all targets, a similar or better candidate was selected when using the quality score of the GNN or GNN-AF ensembles resulting in an increase of 12% (0.615 to 0.69) and 9.43% (0.615 to 0.673), respectively, in the mean DockQ of top 1 selections over the different targets in the benchmark data set compared to AF-M. Classifying the top 1 selected models according to the CAPRI quality categories, the GNN based methods further completely avoids selection of “Incorrect” candidates (Figure 5). Further, using the GNN-AF consensus score, we observe an increase in the proportion of “Medium” ($0.49 \leq \text{DockQ} < 0.80$) and “Acceptable” ($0.23 \leq \text{DockQ} < 0.49$) quality structures. Based on these results, we opt to use the GNN-AF score for the subsequent analyses.

Prediction of TCR-pMHC binding

Next, we turned to the challenge of predicting the correct TCR binding to a given pMHC target. Given the capabilities of the GNN-AF score for ranking docking poses, we hypothesized that it could also be used to separate cognate TCR-pMHC complexes from complexes with swapped incorrect pairings. The intuition being that swapped complexes to a lesser degree would resemble “real” binding complexes, and therefore would tend to be scored lower by quality evaluation metrics.

To investigate this, a TCR specificity dataset consisting of binding and swapped TCR and pMHC was created as described

in the materials and methods section. Briefly, the dataset was generated by downsampling the TCR specificity dataset created in Jensen (2). The downsampling was made to ensure a more even distribution in the number of complexes for each peptide, and to allow for sampling of a large number of structural models for each complex, which would otherwise not be possible for the complete dataset due to computational complexity. The resulting data set consists of 2,945 binding complexes each matched with 5 swapped negative (i.e. the pMHC matched with TCRs positive to other pMHCs) complexes, resulting in $14,725 = 5 * 2,945$ negatives giving a total of 17,670 examples spanning 26 peptides (for more details, see materials and methods).

The different methods were next evaluated in a batch setup, where each batch contained one positive and 5 swapped negative complexes. Structures for each complex were modeled as described in the materials and methods section and the top 1 candidate for each complex selected using a given scoring approach. Next, the complexes within the batch were scored and a metric termed TPR (for details see materials and methods) was computed, which quantifies how many of the negative (swapped) complexes in the batch were scored higher than the positive example. A TPR of 1.0 thus indicates that the positive example was assigned the highest score in the batch. We then sorted batches by the highest intra-batch score predicted by a scoring method in order to compute a cumulative TPR curve for all batches (Figure 6a). Here, we further included a variant of the GNN method trained including the ESM-IF1 inverse folding embedding as input for each residue in the network. This was done to investigate if such representations could serve to boost the method's ability to discriminate the binding interface between true and swapped complexes. When evaluated in terms of top 1 candidate selection and correlation between predicted quality and DockQ values, this method demonstrated a comparable performance to that of the GNN method, both alone and in combination with AF-M (see Table 1).

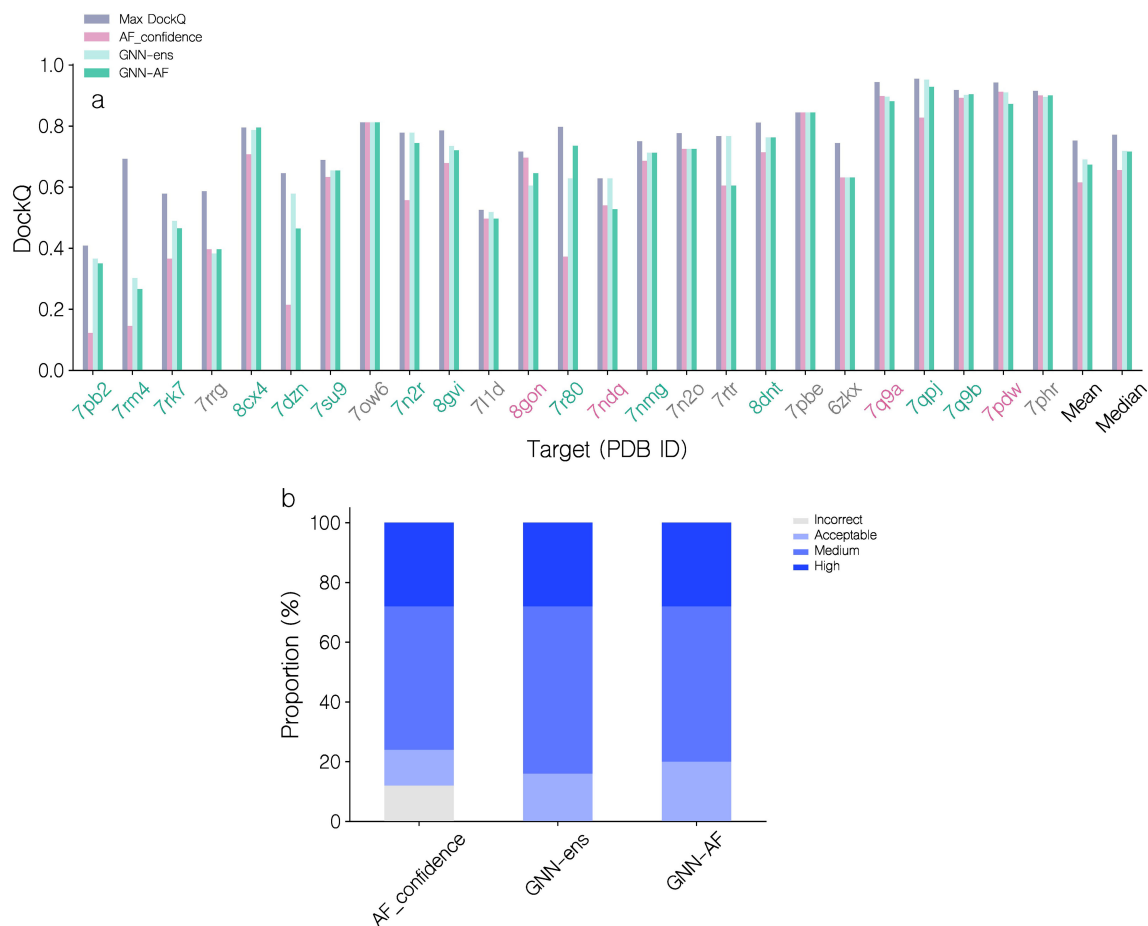


FIGURE 5

Performance evaluation of AF-M, GNN and GNN-AF quality prediction methods for top 1 candidate selection. **(a)** DockQ of top 1 selections for each target, using AF_confidence, GNN-ens or GNN-AF scoring. DockQ of the highest quality candidate is also shown for each target. Red PDB ID label: AF_confidence results in a better selection, gray: identical selection, green: GVP-GNN based scoring results in a better selection. Targets are sorted by their top 1 GNN-AF score. **(b)** Proportions of CAPRI quality categories for top 1 selections for the three methods.

Figure 6 demonstrates that methods combining GNN and AlphaFold-M scoring metrics tend to have a greater cumulative TPR curve AUC, specifically for the first 1000 batches, where GNN-IF1-AF achieves an AUC of 0.76, compared to AUCs of 0.71, 0.72 and 0.69 for AF_confidence, CDRpep_pLDDT and GNN-ens-IF1 respectively. By way of example, the GNN-IF1-AF ensemble maintains a cumulative TPR above 0.8 (corresponding to an average rank of top 2 within the batch size 6) for the top ~300 (~10.2%) batches. This suggests that when a batch contains sufficiently accurate structural models, this consensus method can accurately separate binding complexes from non-binders. In contrast, the individual scoring methods, AF_confidence, AF-CDRpep_pLDDT, GNN-ens-IF1 and particularly GNN-ens archives this high accuracy for a much smaller set of batches. Here, the drop below 0.8 in cumulative TPR occurs already at 60 (2%) and 110 (3.7%) batches for AF_confidence and AF-CDRpep_pLDDT respectively. Notably, the GNN-ens score performed worse in selecting binding complexes compared to the other methods, despite its superior performance in ranking and selecting high modeling quality docking poses (Table 1). It is not

clear what is driving this drop in performance. Therefore, from these results, it is evident that combining the AlphaFold and GNN-ens-IF1 scoring methods into a consensus score, yields superior predictive power for selecting binding TCR-pMHC complexes. Further examining the TPR as a function of the maximal batch scores of the various scoring methods allows us to identify a threshold value at which we can expect accurate ranking within a batch (Figure 6b). Focusing on the consensus GNN-IF1-AF score, we see that the average accumulated TPR of 0.8 corresponds to a predicted quality score of approximately 0.825.

Plotting the distribution of peptides observed for each of batches sorted by the GNN-AF-quality score, we observe that high scoring batches are primarily observed for the peptides LLW and ELA (Figure 6c). That is of the top 100 batches, 79.2% are for the peptides LLW and ELA. This distribution is however more spread out within the top 300 batches (corresponding to the point where the cumulative TPR falls below 0.8). Here, a total of 7 peptides each contribute more than 5% to the distribution.

In conclusion, these results suggest that the proposed quality assessment scoring scheme can accurately separate correct from

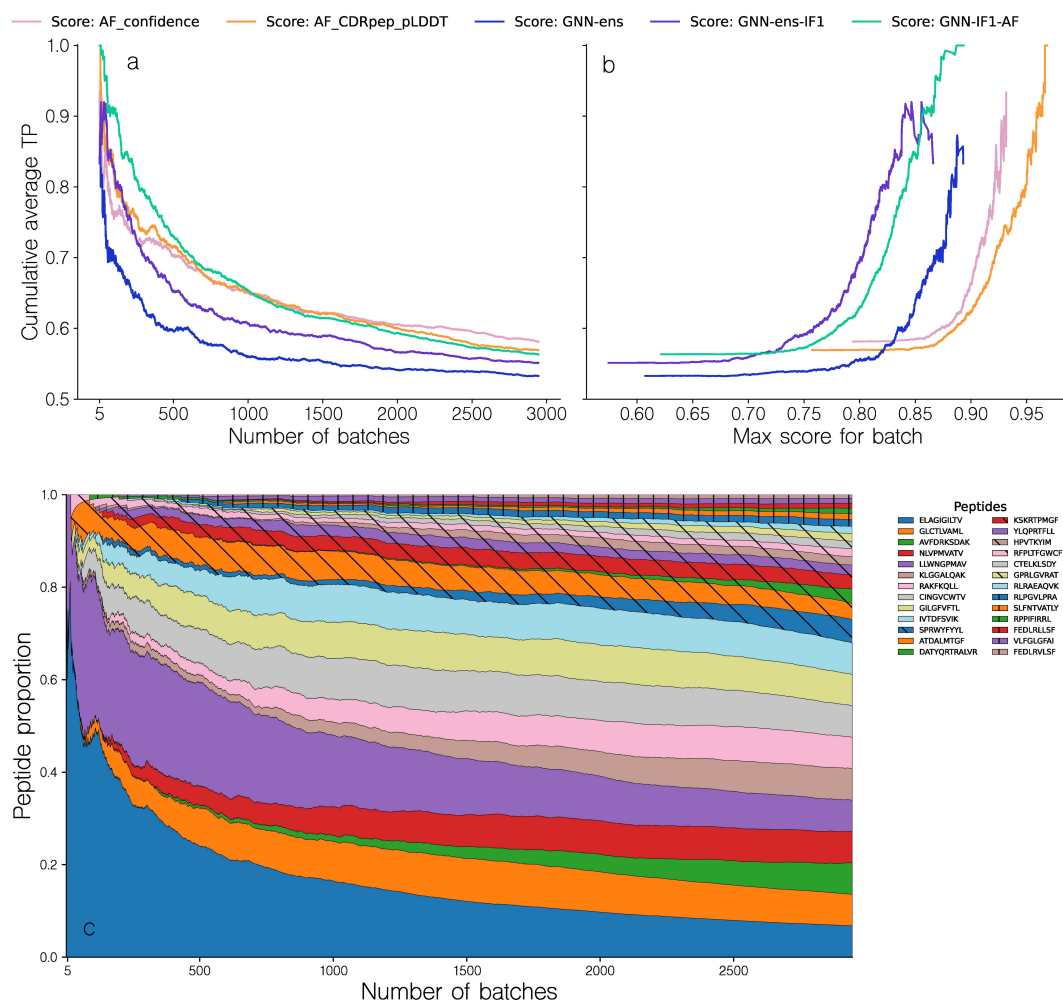


FIGURE 6

Performance evaluation of the different methods in the batch evaluation. **(a)** Cumulative average TPR curve for batches sorted by descending maximal intra-batch quality score for a given method. The TPR value for a given batch is computed from the ranking of the binding complex within each batch, based on their predicted quality, with a value of 1 corresponding to a top 1 rank. **(b)** Cumulative average TPR curve for batches as a function of the maximal intra-batch quality score. The different methods are, AF_confidence: AlphaFold-M confidence, AF_CDRpep_pLDDT: Mean pLDDT of CDR123 and peptide residues, GNN-ens: GVP-GNN ensemble, GNN-ens-IF1: GVP-GNN ensemble trained including ESM-IF1 structural embeddings, GNN-IF1-AF: Harmonic mean of GNN-ens-IF1, AF_confidence and AF_CDRpep_pLDDT. For more information on the GVP-GNN ensembles, see text and materials and methods. **(c)** Proportion of observed peptides in a range of batches sorted according to their max intra-batch score. Note that to avoid showing averages over small numbers, the first data point shown in all plots corresponds to the top 5 batches.

wrongly matched TCR-pMHC paired in cases where the predicted structural accuracy is high. The results however also indicate that is the case in relatively few instances.

Structural inaccuracies may obfuscate binder signal

Given these results and the observation that few TCR-pMHC complexes were modeled with sufficiently high quality to allow for reliable target assessment, we hypothesized that the structure pipeline was unable to produce structural models with sufficiently accurately docked TCRs. To further investigate this, we first recreated the binding classification for models from the structure modeling benchmark dataset (for details refer to materials and

methods). However, given the low number of structures, we opted here not to use homology reduction on this set, giving us 38 structures as opposed to the 25 as described above. For each of the 38 complexes, now denoted as the “binding” complexes, 5 non-binding complexes were generated by swapping the TCR with TCRs randomly selected from other complexes. To avoid sampling potential cross-reactive TCRs as negatives, for a given complex TCRs were only sampled from complexes with a peptide Levenshtein distance to the original complex of at least 3. This resulted in a dataset of 228 complexes. Following this, top 1 candidate structures were selected for each complex using the various scoring methods, and the batch sampling experiment was repeated (Figure 7). For this dataset, the AF confidence metric struggles to correctly rank binding complexes, achieving average cumulative TPR of only 0.5 for the first 20 batches, equivalent to

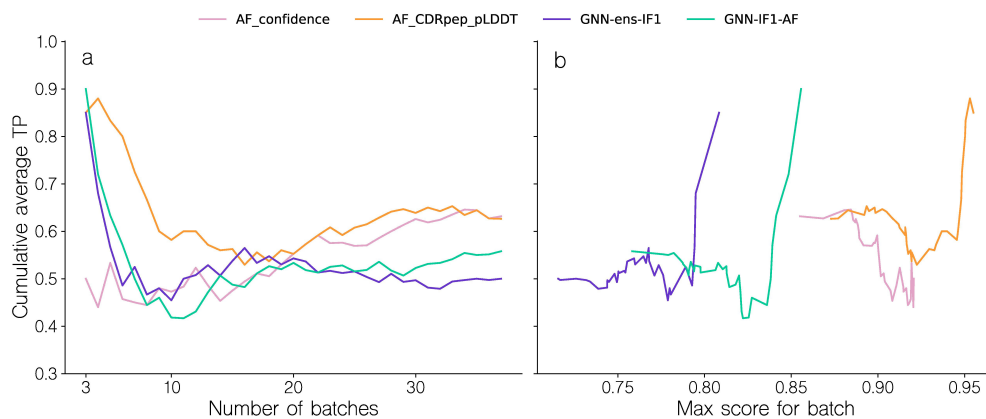


FIGURE 7

Batch sampling TPR curves for benchmark dataset models. **(a)** Cumulative average TPR curve for batches sorted by their TPR. The TPR values are computed by how accurately binding and swapped complexes are ranked within each batch, based on their predicted quality. A TPR closer to 1.0 indicates more correct ranking within batches. **(b)** Cumulative average TPR curve for batches sorted by the max intra-batch quality score. Note that to avoid showing averages over small numbers, the first data point shown in all plots corresponds to the top 3 batches.

random ranking. Notably, the AF_CDRpep_pLDDT score achieves the highest TPR AUC out of all methods on this dataset, maintaining an average cumulative TPR above 0.6 for the first 10 batches.

The GNN based methods, GNN-ens-IF1 and GNN-IF1-AF archive high TPR values only for the first 5 (13%) batches, but subsequently quickly fail to produce accurate rankings. However, the drop in cumulative TPR happens around the same quality_score threshold of 0.825 as observed in Figure 6, corroborating the idea that we can use the maximal intra-batch quality score to determine which batches can be accurately ranked. These results, thus overall align with the finding from the larger-scale NetTCR binding data set of Figure 6. This means, we potentially from this structural data set can investigate properties related to modeling quality, and point to sources for their low accuracy.

For this, we in Figure 8 first display DockQ and the predicted GNN-IF1-AF quality scores of the complexes generated for the positive TCR-pMHC examples as a function of the mini-batch TPR. Here, we can observe significant correlation coefficients between modeling quality measures and classification success ($r=0.65$, p -value $=9.8 \cdot 10^{-6}$, two-tailed t -test test). Focusing on the aggregate DockQ quality measure, we observe that moderately high model quality ($\gtrsim 0.6$ DockQ) is a needed but not sufficient prerequisite for achieving batch TPR values over 0.5. Further, notably, only batches with very high quality positives ($\gtrsim 0.8$ DockQ) are all ranked very accurately, with most batches achieving a TPR of 1.0. This is corroborated by the GNN-IF1-AF score, which is shown to be more strongly correlated with TPR ($r=0.88$, p -value $=1.7 \cdot 10^{-13}$, two-tailed t -test test), with batches with a positive complex with a score above 0.8 all archiving a TPR value 0.8 or above. These results

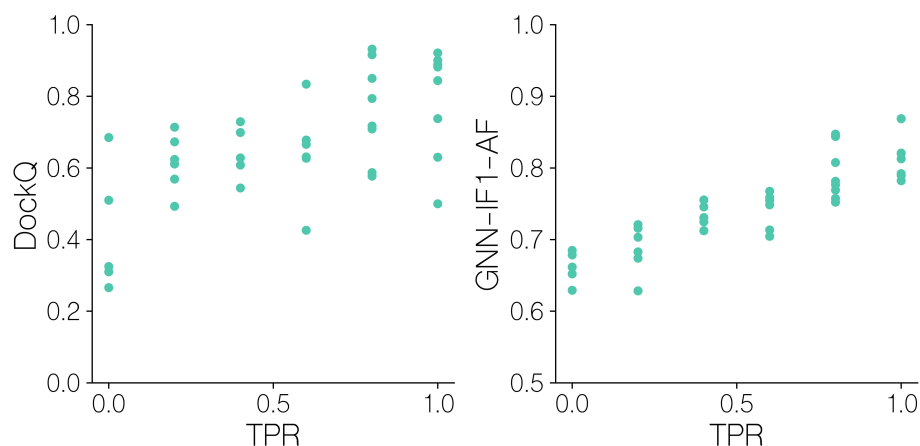


FIGURE 8

Batch sampling TPR values versus quality metrics for benchmark dataset models. DockQ and GNN-IF1-AF scores versus mini-batch ranking TPR for models of solved structures. GNN-IF1-AF: combined GVP-GNN-IF1 ensemble, AF-M confidence and peptide-CDR pLDDT scores (see materials and methods). A TPR of 1.0 indicates that in the given batch, the positive example was assigned the highest GNN-IF1-AF score.

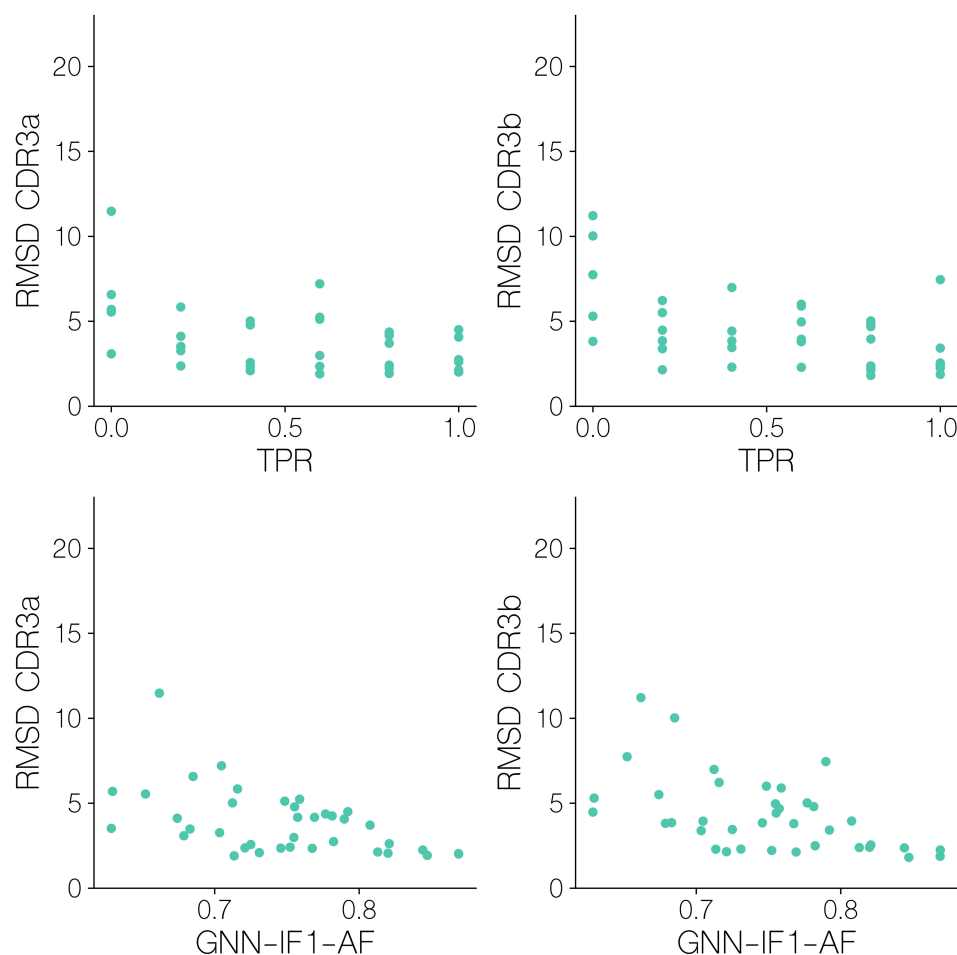


FIGURE 9
CDR3 $\alpha\beta$ backbone/sidechain RMSD after superimposition on pMHC C α atoms for benchmark set targets, versus their GNN-IF1-AF score and mini batch TPR.

show that structural modeling quality clearly influences TCR-pMHC binding classification success. Particularly, that high quality models ($\gtrsim 0.8$ DockQ and $\gtrsim 0.8$ GNN-IF1-AF score) are required for accurate classification. To further this analysis, we computed CDR3ab RMSDs after superimposing the models onto the pMHC and examined their relationships with predicted quality and TPR (Figure 9).

Here, we observe that for both CDR3a and CDR3b backbone/sidechain atoms, an RMSD of $< 2.5\text{\AA}$ is required in order to reach the aforementioned GNN-IF1-AF score threshold of ~ 0.8 that predicts classification success, that we see very few structural models reach. We see that mini-batch TPR is significantly correlated with both CDR3a and CDR3b RMSD ($r = -0.462$, $p\text{-value} = 3.52 \cdot 10^{-3}$, $r = -0.529$, $p\text{-value} = 6.43 \cdot 10^{-4}$, two-tailed t-test), showing that docking accuracy significantly influences how accurately we can assess binding and non-binding complexes. Particularly for CDR3b loop, with the exception of an outlier, all batches with a TPR of 1 have an RMSD of $< 2.5\text{\AA}$. This again, suggests a modeling quality threshold that is difficult to attain when also considering findings from the modeling quality benchmark (Figure 5) and the observed classification performance (Figure 6).

Predictive performance on unseen peptides in IMMREP23

In order to further validate the predictive performance of the GNN-IF1-AF score, we evaluated the model on a subset of the IMMREP23 TCR specificity prediction benchmark dataset (9). Specifically, we focused on data for the previously unseen peptides SALPTNADLY, TSDACMMTMY and FTDALGIDEY. We modeled the 246 TCR-pMHC complexes from this subset using the modeling pipeline described here and scored the resulting models using a set of different scoring methods including AF-M, GNN and the GNN-IF1-AF ensemble scores. However, for this dataset we opted to only model 10 candidates per complex and without any restrictions on template selection in order to mimic a more realistic use case of the modeling and scoring pipeline. We then selected the top 1 scoring complex, using each of the scoring methods, and used this score to compute an AUC 0.1 value for each peptide (Figure 10).

On this benchmark dataset, we find that the AlphaFold-M scores, AF_confidence and AF_CDRpep_pLDDT achieved a slightly higher AUC0.1 than the GNN based scores. This is the

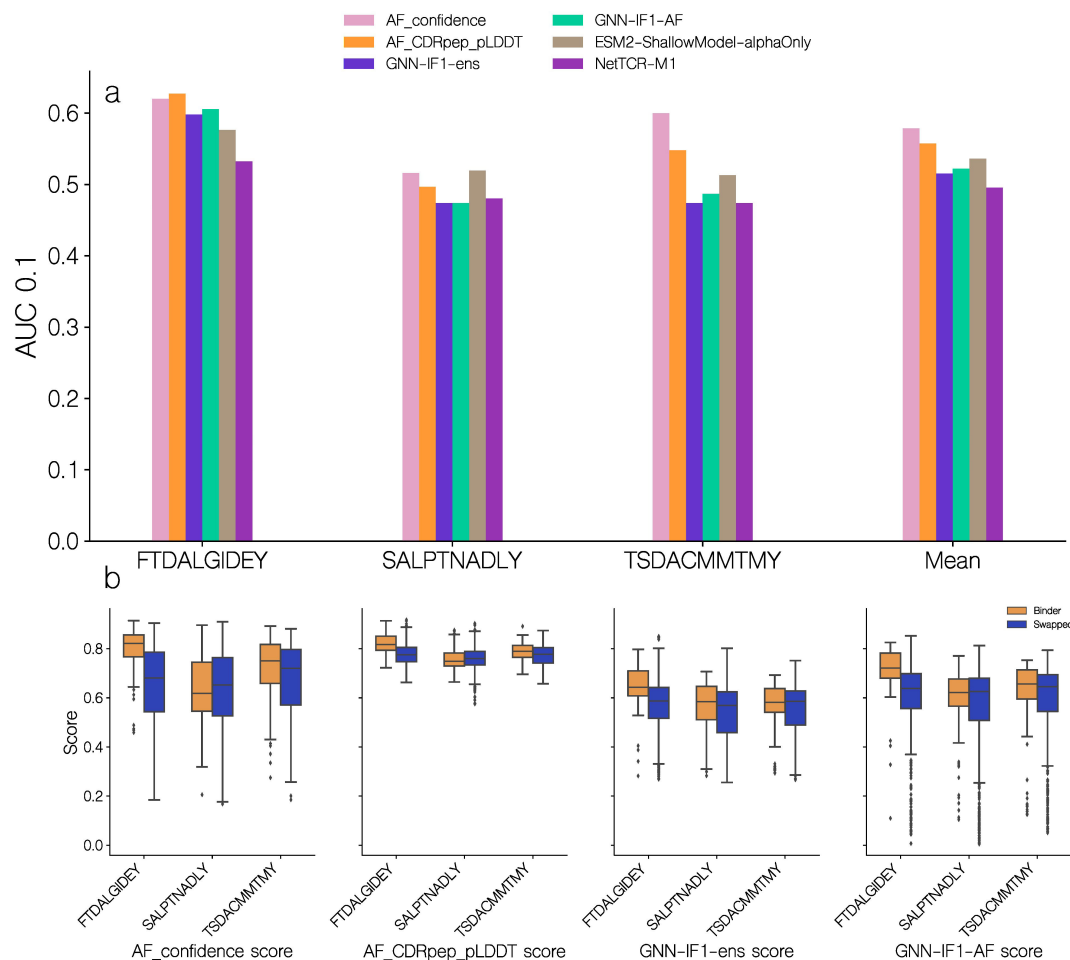


FIGURE 10

Predictive performance on a subset of the IMMREP23. **(a)** Predictive performance of quality scoring methods on data for the unseen peptides in the IMMREP23 benchmark dataset. The AUC 0.1 is computed from the predicted quality score, obtained from the top 1 score of each scoring method. AF_confidence: AlphaFold-M confidence, AF_CDRpep_pLDDT: Mean pLDDT of CDR123 and peptide residues, GNN-ens: GVP-GNN ensemble, GNN-ens-IF1: GVP-GNN ensemble trained including ESM-IF1 structural embeddings. GNN-IF1-AF: Harmonic mean of GNN-ens-IF1, AF_confidence and AF_CDRpep_pLDDT. For more information on the GVP-GNN ensembles, see text and materials and methods. ESM2-ShallowModel: A 3-layer multilayer perceptron train on TCR and peptide protein sequence embeddings from ESM2. NetTCR-M1: 1 dimensional convolutional neural network trained on BLOSUM50 embeddings of TCR and peptide sequences (27). **(b)** Distribution of predicted quality scores for each of the unseen IMMREP23 benchmark peptides.

case for all three peptides, however mostly pronounced for TSD, where the AF_confidence and AF_CDRpep_pLDDT scores achieve AUC0.1 values of 0.60 and 0.548 respectively, while the GNN based methods, GNN-IF1-ens and GNN-IF1-AF, achieve values of 0.474 and 0.487 respectively. For the remaining two peptides, all methods perform similarly. On SAL, performance is poor across all methods, with AUCs around 0.5, while on FTD, they all achieve AUCs of approximately 0.6 indicating better than random classification. Plotting the distribution of scores predicted by the different methods for each peptide (Figure 10b), we find all methods share higher scoring values for the FTD peptides, which the methods can more accurately predict binding for. For the GNN-IF1-AF method in particular, only FTD has complexes with scores >0.8, where 12% and 2% of binders and non-binders respectively surpass this threshold. The proportion of complexes with a score >0.7, for each peptide for binders and non-binders are 61% and 24%, 15%

and 19% and 34% and 22% for peptides FTD, SAL and TSD respectively. Thus, we see an enrichment in higher quality structural models for binders, for 2 out of 3 peptides. These observations corroborate our hypothesis, that only in cases with high quality structural models, we can use these to predict binding. We also here compared the performance of these scoring methods to the performance of two models that were entered in the IMMREP23 benchmark. Both of these models use amino acid sequence data to generate predictions, where one is a multilayer perceptron taking ESM-2 protein language model embeddings as input and the other is a convolutional neural network taking BLOSUM50 embeddings as input (27, 28). Here, we find that all methods based on structural data outperform the sequence based models on FTD, and the AlphaFold-M scores for TSD. While the methods utilizing GNN ensembles in this particular benchmark only achieved similar performance to the sequence based models,

these results demonstrate that structural data can assist in predicting TCR-pMHC binding for novel peptides.

NetTCR-struc GitHub repository

We have made our GNN-AF and GNN-IF1-AF methods available in a GitHub repository for use in docking scoring and TCR-pMHC binding prediction at <https://github.com/mnielLab/NetTCR-struc>.

Discussion

In this work, we made efforts to improve structural modeling accuracy of TCR-pMHC class I complexes and evaluated the use of structural data for predicting TCR-pMHC binding. For these tasks, we evaluated a range of methods, including AlphaFold-Multimer's internal scoring metrics as well as graph neural network based methods trained on structural models of TCR-pMHC complexes.

Structural diversity and modeling accuracy

Initially, we observed that AF-M produced structurally homogeneous docking candidates, often misrepresenting the CDR loop configurations. Using MSA feature perturbation, we demonstrated that we could generally improve modeling diversity, however this did not lead to any meaningful improvements in overall modeling quality. However, given its mixed effects on modeling accuracy, we opted to use this approach selectively for training a scoring function rather than as a default setting.

A major challenge in model selection was the overestimation of model quality by AF-M's internal confidence metric. Our analysis revealed only a moderate correlation between AF-M confidence scores and DockQ scores, leading to incorrect assessment of low-quality models. To improve this, we developed a series of GNN-based scoring methods trained to predict DockQ scores. Our best-performing model, GNN-AF, combined GNN predictions with AF-M confidence metrics, yielding a 21% increase in Spearman's correlation (from 0.681 to 0.855) with DockQ, significantly improving ranking accuracy for docking candidates.

Touching upon DockQ as a metric for evaluating TCR-pMHC modeling quality, we noted that the DockQ values reported for the complexes in the benchmark set appeared to be generally high with a mean of 0.673 DockQ for the top 1 selections. However, due to the highly conserved nature of TCR-pMHC docking geometry these numbers are in fact an exaggeration of the actual modeling skill of AlphaFold-M. Thus, while some targets such as 7PHR, 7Q9B, 7PWD, 7Q9A and 7QPJ that are modeled with near-native quality likely do indeed capture most pMHC and TCR interactions, models for the remaining targets, likely fail to

capture important interactions in the docking interfaces, despite their generally high DockQ values.

Prediction of TCR-pMHC binding

Using the methods developed in the docking pose quality scoring, we explored the potential for using modeling quality as a score for TCR-pMHC binding prediction. Here, our results demonstrated that integrating GNN-AF with inverse folding embeddings (GNN-IF1-AF) further improved discrimination of binding versus non-binding complexes, particularly when high-quality structural models were available. However, while the GNN-IF1-AF ensemble outperformed AF-M scoring metrics, it (along with all other methods) only maintained a high performance for a small proportion of analyzed data. This therefore underlined that the structural pipeline in general struggled to generate sufficiently high-quality TCR-pMHC models suitable for binding prediction. This was further elucidated when evaluating binding classification performance on the solved structures docking pose ranking benchmark dataset. Here, the AF confidence metric failed completely to effectively rank binding complexes, whereas the GNN-IF1-AF method performed better for a small subset of high quality structures but otherwise failed. This again suggests that structural modeling quality is a key determinant of ranking performance.

While conducting this research, AlphaFold-3 was released which demonstrates significant improvements in modeling quality, particularly for antibodies. Due to the related nature of TCRs and antibodies, we might expect this version to also showcase improvements in modeling for TCR-pMHC complexes over AlphaFold-2.3. However, since the docking scoring approach we have developed here is orthogonal to the AlphaFold scoring scheme, we still expect the approach of applying (a potentially refined) TCR-pHLA specific quality scoring model would help improve ranking of structural models created also using AlphaFold-3. The same argument holds when it comes to our conclusions on the limitations of structural modeling and how it affects TCR-pMHC binding. While the newest structural modeling approaches might improve on this, we still expect the challenge to remain for parts of the TCR-pMHC binding space. To access and characterize this, we recommend further investigations are conducted along the lines of the work described here.

In conclusion, we have here presented a pipeline incorporating inverse-folding embeddings and GNN-based scoring for refined quality assessment of TCR-pMHC structures. This approach was found to significantly improve structural ranking and binding prediction, and further provided insights into how structural modeling quality affects binding prediction performance. However, our findings also elucidated key limitations in current structural modeling methods. The release of AlphaFold-3 may offer improvements in modeling quality, thereby potentially addressing some of the observed shortcomings. However, our results suggest that domain specific, novel quality scoring approaches like those

present here will remain useful for improving on their corresponding structural modeling tasks. Continued work on this topic will allow limitations both in terms of the structural modeling and scoring accuracies to be identified and characterized, enabling new pathways for refining docking scoring approaches and subsequently improving the accuracy and reliability of TCR-pMHC binding predictions.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://github.com/mnielLab/NetTCR-struct>.

Author contributions

SD: Visualization, Data curation, Formal analysis, Methodology, Writing – review & editing, Investigation, Software, Conceptualization, Writing – original draft. MN: Supervision, Writing – review & editing, Project administration, Visualization, Formal analysis, Methodology, Investigation, Funding acquisition, Conceptualization, Writing – original draft, Resources.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. Research reported in this publication was supported in part by the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 101007799 (Inno4Vac). Research reported in this publication was supported in part by the National Institute of Allergy and Infectious Diseases (NIAID), under award number 75N93019C00001.

References

1. Townsend A, Bodmer H. Antigen recognition by class I-restricted T lymphocytes. *Annu Rev Immunol.* (1989) 7:601–24. doi: 10.1146/annurev.iy.07.040189.003125
2. Jensen MF, Nielsen M. Enhancing TCR specificity predictions by combined pan- and peptide-specific training, loss-scaling, and sequence similarity integration. *eLife.* (2024) 12:RP93934. doi: 10.7554/eLife.93934
3. Meynard-Piganeau B, Feinauer C, Weigt M, Walczak AM, Mora T. TULIP: A transformer-based unsupervised language model for interacting peptides and T cell receptors that generalizes to unseen epitopes. *Proc Natl Acad Sci U. S. A.* (2024) 121:e2316401121. doi: 10.1073/pnas.2316401121
4. Moris P, De Pauw J, Postovskaya A, Gielis S, De Neuter N, Bittremieux W, et al. Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Brief Bioinform.* (2021) 22:bbaa318. doi: 10.1093/bib/bbaa318
5. Han Y, Yang Y, Tian Y, Fattah FJ, von Itzstein MS, Hu Y, et al. pan-MHC and cross-species prediction of T cell receptor-antigen binding. *BioRxiv prepr Serv Biol.* (2023). doi: 10.1101/2023.12.01.569599
6. Croce G, Bobisse S, Moreno DL, Schmidt J, Guillame P, Harari A, et al. Deep learning predictions of TCR-epitope interactions reveal epitope-specific chains in dual alpha T cells. *Nat Commun.* (2024) 15:3211. doi: 10.1038/s41467-024-47461-8
7. Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. *Nature.* (1988) 334:395–402. doi: 10.1038/334395a0
8. Meysman P, Barton J, Bravi B, Cohen-Lavi L, Karnaukhov V, Lilleskov E, et al. Benchmarking solutions to the T-cell receptor epitope prediction problem: IMMREP22 workshop report. *ImmunoInformatics.* (2023) 9:100024. doi: 10.1016/j.immuno.2023.100024
9. Nielsen M, Eugster A, Jensen MF, Goel M, Tiffeau-Mayer A, Pelissier A, et al. Lessons learned from the IMMREP23 TCR-epitope prediction challenge. *ImmunoInformatics.* (2024) 16:100045. doi: 10.1016/j.immuno.2024.100045
10. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee JY, et al. Diversity and clonal selection in the human T-cell repertoire. *Proc Natl Acad Sci U. S. A.* (2014) 111:13139–44. doi: 10.1073/pnas.1409155111
11. Coles CH, Mulvaney RM, Malla S, Walker A, Smith KJ, Lloyd A, et al. TCRs with distinct specificity profiles use different binding modes to engage an identical peptide-HLA complex. *J Immunol Baltim Md 1950.* (2020) 204:1943–53. doi: 10.4049/jimmunol.1900915
12. Vita R, Blazeska N, Marrama D, IEDB Curation Team Members, Duesing S, Bennett J, et al. The immune epitope database (IEDB): 2024 update. *Nucleic Acids Res.* (2025) 53:D436–43. doi: 10.1093/nar/gkae1092
13. Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* (2018) 46:D419–27. doi: 10.1093/nar/gkx760
14. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv.* (2022). doi: 10.1101/2021.10.04.463034

Acknowledgments

This manuscript has been posted as a preprint on biorxiv, with doi: <https://doi.org/10.1101/2025.03.22.644721>.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fimmu.2025.1616328/full#supplementary-material>

15. Yin R, Ribeiro-Filho HV, Lin V, Gowthaman R, Cheung M, Pierce BG. TCRmodel2: high-resolution modeling of T cell receptor recognition using deep learning. *Nucleic Acids Res.* (2023) 51: gkad356. doi: 10.1093/nar/gkad356
16. Bradley P. Structure-based prediction of T cell receptor:peptide-MHC interactions. *eLife.* (2023) 12:e82813. doi: 10.7554/eLife.82813
17. Slone JK, Conev A, Rigo MM, Reuben A, Kaviraki LE. TCR-pMHC binding specificity prediction from structure using graph neural networks. *IEEE Trans Comput Biol Bioinforma.* (2025) 22:171–9. doi: 10.1109/TCBBIO.2024.3504235
18. Golovin AV, Pantelev S, Zlobin AS, Anikeeva N, Smirnov I, Gabibov A, et al. The role of peptide conformation presented by MHC in the induction of TCR triggering. *Biophys J.* (2025) 124:1073–84. doi: 10.1016/j.bpj.2025.02.001
19. Motmaen A, Dauparas J, Baek M, Abedi MH, Baker D, Bradley P. Peptide-binding specificity prediction using fine-tuned protein structure prediction networks. *Proc Natl Acad Sci.* (2023) 120:e2216697120. doi: 10.1073/pnas.2216697120
20. Basu S, Wallner B. DockQ: A quality measure for protein-protein docking models. *PLoS One.* (2016) 11:e0161879. doi: 10.1371/journal.pone.0161879
21. Jing B, Eismann S, Suriana P, Townshend R, Dror R. Learning from protein structure with geometric vector perceptrons. *arXiv: arXiv:2009.01411.* (2021). doi: 10.48550/arXiv.2009.01411
22. Stein RA, Mchaourab HS. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with AlphaFold2. *PLoS Comput Biol.* (2022) 18: e1010483. doi: 10.1371/journal.pcbi.1010483
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. *Nucleic Acids Res.* (2000) 28:235–42. doi: 10.1093/nar/28.1.235
24. Hsu C, Verkuil R, Liu J, Lin Z, Hie B, Sercu T, et al. Learning inverse folding from millions of predicted structures. *bioRxiv.* (2022). doi: 10.1101/2022.04.10.487779
25. Povlsen HR, Bentzen AK, Kadivar M, Jessen LE, Hadrup SR, Nielsen M. Improved T cell receptor antigen pairing through data-driven filtering of sequencing information from single cells. *eLife.* (2023) 12:e81810. doi: 10.7554/eLife.81810
26. A new way of exploring immunity - linking highly multiplexed antigen recognition to immune repertoire and phenotype. *Immunol Microbiol Technol Networks.* (2025). Available online at: <http://www.technologynetworks.com/immunology/application-notes/a-new-way-of-exploring-immunity-linking-highly-multiplexed-antigen-recognition-to-immune-repertoire-332554> (Accessed March 21, 2025).
27. Jensen MF, Nielsen M. NetTCR 2.2 - Improved TCR specificity predictions by combining pan- and peptide-specific training strategies, loss-scaling and integration of sequence similarity. *eLife.* (2024) 12:RP93934. doi: 10.7554/eLife.93934.2
28. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv.* (2022). doi: 10.1101/2022.07.20.500902
29. Simister PC, Border EC, Vieira JF, Pumphrey NJ. Structural insights into engineering a T-cell receptor targeting MAGE-A10 with higher affinity and specificity for cancer immunotherapy. *J Immunotherapy Cancer.* (2022) 10(7). doi: 10.1136/jitc-2022-004600